Creating a Semantically-Annotated Corpus of Tweets in Catalan

Blanca Calvo Figueras **Mar Rodríguez Álvarez** **Celia Soler Uguet**

Abstract

Semantic information has been shown to be relevant in many NLP tasks. However, good semantic parsers can only be developed if there exist enough semantically-annotated data. This is often not the case for low-resource languages. This project aims at creating a semantically-annotated corpus in Catalan. The corpus is comprised of 50 tweets, which are annotated following the guidelines from PropBank and AnCora on argument annotation. Given the particular characteristics of the dataset, we set additional guidelines for the annotation. We also create new entries of verbs that are not present in the current AnCora lexicon.

1 Introduction

Recent studies have found that current state-of-the-art Language Models (Vaswani et al., 2017; Yang et al., 2020) might not be grasping important aspects of language (Mudrakarta et al., 2018; Zanzotto et al., 2020). Consequently, they fail to perform some tasks that require a deep understanding of language. Neural models encode sentences into embedding representations, which do not consider the various predicate-argument structures used in human language. For this reason, later research has introduced semantic information, such as Semantic Roles, to improve the performance of natural language understanding tasks (Zhang et al., 2020; Zhong et al., 2020).

There currently exist semantic role labelers in English that have a satisfactory performance (Shi and Lin, 2019; Brown et al., 2019). However, to train these models for languages with less resources, it is important to have semantically-annotated corpora available. In this project, we contribute to the development of semantic language resources for Catalan.

For this purpose, we annotate a corpus of text in Catalan using PropBank semantic roles (Palmer et al., 2005). [1] We chose PropBank over VerbNet because the earlier is the most popularly used in NLP. We base our annotations in the AnCora_CA corpus (Taulé et al., 2011). This is a semantically-annotated corpus, which considers arguments and semantic roles. AnCora_CA also includes a lexicon of Catalan verbs with its arguments.

To create our corpus we will be using data from the social network Twitter. We decided to do so, because it is a source of current colloquial written language, which can easily be retrieved through Twitter's API.

The goal of this language resource is two-fold. (1) Creating an annotated corpus that can be used in the future to develop automatic systems for semantic parsing. (2) The systematic and organised annotation of a corpus can contribute to discovering semantic structures that are not included in the current lexicon but are being used by social media users.

2 Annotation guidelines

In this section we present the guidelines used in our annotation task. The topics that will be covered are the following:

- what has been annotated,
- which categories have been used for the annotation,
- how the corpus has been annotated.

Our approach towards the annotation task consists in finding the verbs of each sentence in a tweet, annotating the arguments of each verb, and choosing to which argument each of the words of the

[1] The corpus can be found in https://github.com/BlancaCalvo/Tweet-Sem-Cat

sentence belong. In order to perform this task, we create a simple interface (Section 2.3).

Given that our corpus is extracted from Twitter, the particularities of the data have to be taken into account. We have included some additional guidelines to deal with noise in our corpus: such as tweets written in a language other than Catalan, tweets without a verb and non-comprehensible tweets (Section 2.2). Additionally, punctuation marks belonging to the arguments have been annotated as such. However, punctuation marks at the end of sentences between arguments have not been annotated.

In the event of finding a verb that is not contained in AnCora, we create entries and use them for the annotation (Section 6).

2.1 AnCora Guidelines

For this task, we will follow the guidelines from AnCora for argument annotation in Catalan (Taulé et al., 2011).

Four general semantic classes are specified in the corpus: states, activities, accomplishments and achievements. They can be subclassified depending on thematic roles and diatheses which represent the correspondence between semantic valence and syntactic actant in lexical units of predicator. However, what we will cover in our project is the annotation of verb arguments.

In this matter, AnCora follows the proposal from PropBank (Palmer et al., 2005), distinguishing seven possible argumentals: arg0, arg1, arg2, arg3, arg4,argM and argL.

The first five arguments are internal arguments and they are ordered regarding their relation with the verb, being arg0 the most related argument to the verb, and arg4 the least one. ArgM corresponds to adjuncts and argL to lexicalised complements of light verbs.

Each argument is explained below:

- Arg0: external causer of the verb; given to the subject of transitive and inergative verbs and to the agent complement of passives.
- Arg1: first internal argument of the verb; given to the direct object of transitive verbs and to the subject of unaccusatives and statives.
- Arg2: second internal argument of the verb; given to the indirect or prepositional objects of ditransitive verbs and to attributes or prepositional objects of statives and unaccusatives.
- Arg3: starting point of change of place or change of state predicates
- Arg4: ending point of change of place or change of state predicates.
- ArgM: non-argument constituents (adjuncts) in all verb classes.
- ArgL: lexicalized complements of lights verbs which admit variation to some extent.

Moreover, AnCora differenciates 20 different thematic roles and 12 functions. These will not be used for the annotation, but will be taken into account when searching for the most accurate argument in AnCora. All the correspondences between arguments, thematic roles and functions are explained in Annex A.

2.2 Additional Guidelines

Since our corpus comes from Twitter's API, not all our data is clean and useful. To deal with these cases, we have created additional annotation guidelines for the task:

- NC: Non-comprehensible tweet.
- NV: Tweet without verb.
- OL: Tweet written in another language.

In table 1 we show some examples of tweets that have been annotated following these guidelines.

2.3 Annotation Interface

For the annotation task, we have create a Python program that helps us annotate faster. The program shows us the tweet and then, we state whether we will proceed with the annotation or not by using the labels 'NC', 'NV' or 'OL', that were described above. Then, we indicate where the verb is located and which arguments does it have. And finally, we choose which words belong to each argument. The program, then, outputs an XML file containing the final annotated tweets. In Annex B, an example of the final XML file can be observed. This XML follows the structure of the AnCora corpus in order to make it easy to combine our corpus with AnCora.

Annot.	Examples
Arg0	**Lavia Valeriana** als 12 anys ajudava a la besàvia amb la ràdio a la Résistance i fins ara hackejant al feixisme.
Arg1	Si us agrada **el ballet clàssic**, diumenge a les 11.15h no us perdeu LA VENTAFOCS!
Arg2	Volem donar, de part de tot el Club Voleibol Manacor, les condolences **a família i aficionats del USER** per la recent mort del seu President i Fundador Miquel Jaume Roig.
Arg3	Jo sortint **de la feina** i entrant a twitter per primera vegada en tot el dia URL
Arg4	Langlès arriba **a Catalunya Ràdio**.
ArgM	Ets dels que està convençut que a la parella se la sedueix **per l'estómac**?
ArgL	Amb el ramal de la MAT a la Selva perdrem qualitat de vida i es posarà **en perill** la salut.
NC	a,,, morí URL
NV	La riera de Beget a la palanca del Samsó. #rieradebeget #palancadelsamsó #valldhortmoier #hortmoier #altagarrotxa #garrotxa #oix #bicicleta #gravel #hivern2021 URL
OL	Bhuaigh Abderrahmane El Mahmi an comórtas filíochta sa 5ú bliain. An excellent study resource for 5th and 6th year students preparing for their poetry reading in the oral. Thar barr! #SnaG21

Table 1: Examples of annotations.

3 Data Statement

Following the research on mitigating biases in NLP (Bender and Friedman, 2018), in this section we will give an in-depth explanation of the characteristics of our resource, being those (1) curation rationale, (2) language variety and register, (3) speaker demographic and (4) annotator demographic.

Curation rationale refers to which texts are included in the resource and which were the goals in selecting those texts. In that sense, we chose to use tweets because they are a close representation of the current use of language in social media.

With regard to the language variety, the register contained in the tweets tends to be colloquial. In that sense, spelling and punctuation mistakes are common. Moreover, as far as register is concerned, since the selection of tweets is random, it will contain different dialects of the Catalan language.

Regarding the speaker demographic, since the tweets are written in Catalan, we can say that the group represented is Catalan speakers. However, some relevant information is missing, such as the age, gender and race of the speakers. Whether Catalan is their first language or not is also unknown. Such missing information can be crucial for avoiding biases in tasks such as sentiment analysis, which can lead to real consequences in society. However, for the specific task of semantic annotation, we consider this information to have little relevance.

Finally, regarding the demographic characteristics of the annotators, there are several factors to highlight. The annotator's group is conformed of three people. They are all the same gender, same race and around the same age. However, they come from different academic backgrounds (linguistics, translation and digital humanities), and only two of them have Catalan as a mother tongue.

4 ITA agreement

Before annotating the whole corpus, we have annotated the first 10 tweets of the dataset in order to check if our guidelines were clear to all the annotators. We have calculated the Inter Annotator Agreement with the Fleiss Kappa measure. We have obtained the following results:

- IAA for verb detection: 0.764
- IAA for verb sense annotation: 0.912
- IAA for argument annotation: 0.785
- IAA for words in arguments detection: 0.946

Since all the results are between substantial and perfect, we have assessed that our guidelines are good.

5 License

We decided to release our corpus under the license Creative Commons Zero v1.0 Universal, so that it can be freely used for as many people as possible. This license states the following:

"The Creative Commons CC0 Public Domain Dedication waives copyright interest in a work you've created and dedicates it to the world-wide public domain. Use CC0 to opt out of copyright entirely and ensure your work has the widest reach. As with the Unlicense and typical software licenses, CC0 disclaims warranties. CC0 is very similar to the Unlicense."

6 New Entries for AnCora

During the annotation process we expected to find verbs or verb senses that were not included in AnCora. In these cases, we have created new lexicon entries following the entries of the AnCora lexicon. We have not found any verb that was not yet in AnCora; however, we have encountered some senses or constructions which were not contemplated. Those were the following:

- 'deixar': it literally means 'to leave', but in the context of the tweet, it was referring to the figurative meaning of 'dying'.
- 'vetllar': it has a similar meaning to 'safeguard' and AnCora only contemplated the use of arg0 and arg1 with the verb. However, one of our tweets needed to have an argM, which was not included in AnCora.
- 'escriure': it means 'to write'; however, as in the last case, AnCora did not include an entrance containing an arg2 (beneficiary). Therefore, we created a new one.
- 'gaudir': it means 'to enjoy'. In this case, again, AnCora did not contemplate the use of argM with such verb, so we needed to create a new entrance to include this argument.

The full new entries can be found in Annex C.

7 Discussion and Future Work

Having finished the project, we can highlight some work that could be done in the future to improve our resource.

The annotation interface we created tokenizes the tweets automatically. This results in some inaccuracies. For instance, "d'altres", meaning "some others", should have been split into two tokens "de" and "altres". That is also the case of "digue-li", meaning "tell him/her". In this project we have annotated these words together. However, future work could use tokeniser that is specific for Catalan, such as the one of Freeling (Padró and Stanilovsky, 2012).

Further work would also include annotating more tweets in order to provide a large annotated corpus that could be used to create a semantic parser in Catalan.

References

Emily M. Bender and B. Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. Verbnet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question?

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Mariona Taulé, M Antònia Martí, and Oriol Borrega. 2011. Ancora 2.0: Argument structure guidelines for catalan and spanish.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. ArXiv: 1906.08237.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Annex A: Correspondence between Arguments, Roles and Functions

A Correspondences between Arguments, $\theta-$Roles and Functions

Attribute <func> **value** / *function*	**Attribute <arg>** **value**	**Attribute <tem>** **value** / *thematic role*	**Example**
suj / *subject*	**arg0**	**agt** / *agent*	**Juan** lee una novela
		cau / *cause*	**El viento** abrió la puerta
		exp / *experiencer*	**Juan** sueña
		src / *source*	**Juan** sudaba
	arg1	**pat** / *patient*	**Clara** es amada por todos
		tem / *theme*	**Juan** llegó tarde
	arg2	**ins** / *instrument*	**Una lona** cubre el coche de Juan
		loc / *locative*	**El diario** abordó la noticia
	argL	*none*	**El cadáver** fue levantado a la 1 de la tarde
cd / *direct object*	**arg1**	**pat** / *patient*	Juan lee **una novela**
		tem / *theme*	El viento abrió **la puerta**
	arg2	**atr** / *attribute*	Juan tiene **un coche blanco**
		ext / *extension*	El paro subió **15.891 personas**
	argL	*none*	Puso **punto final** a la discusión
creg / *prepositional object*	**arg1**	**tem** / *theme*	Clara se rió **del chiste**
		loc / *locative*	Juan intervino **ante la comisión**
	arg2	**tem** / *theme*	Clara sustituyó el vino **por agua**
		loc / *locative*	Juan apoyó la bicicleta **en un árbol**
		ins / *instrument*	Juan va equipado **con un casco**
		atr / *attribute*	Clara joza **de buena salud**
		cot / *cotheme*	Juan conecta el ordenador **a la impresora**
		efi / *final state*	Reconvirtió la habitación **en un estudio**
		ein / *initial state*	Antonio se recupera **de un accidente**
		ext / *extension*	La crisis situó el paro **en 93.278 personas**
	arg3	**ori** / *origin*	El aceite proviene **de las olivas**
	arg4	**des** / *destination*	Juan llevó el coche **al garaje**
	argL	*none*	El niño estalló **en sollozos**
ci / *indirect object*	**arg2**	**ben** / *beneficiary*	Clara **se** lo sugirió **(a Juan)**
		exp / *experiencer*	**(A Clara) le** gusta pasear
	arg3	**ben** / *beneficiary*	**(A Juan) le** salen muy bien las tortillas
		exp / *experiencer*	El chiste **le** pareció divertidísimo
cag / *agent complement*	**arg0**	**agt** / *agent*	Clara es amada **por todos**
cpred / *predicative complement*	**arg2**	**atr** / *attribute*	El niño se llama **Daniel**
	arg3	**atr** / *attribute*	Juan pasó la tarde **sin pensar en Clara**
	argM	**atr** / *attribute*	Suspiró **embelesado**
	argL	*none*	Clara puso **de relieve** su falta de tacto
atr / *attribute*	**arg2**	**atr** / *attribute*	Clara es **abogado**
cc / *adjunct*	**arg1**	**tem** / *theme*	Juan continúa **con sus estudios**
	arg2	**atr** / *attribute*	Clara se pirra **por las patatas fritas**
		cot / *cotheme*	Juan no está emparentado **con Clara**
		efi / *final state*	Clara está inmunizada **contra disgustos**
		ext / *extension*	El juicio se prolongó **hasta el día 15**
		ins / *instrumental*	Juan adornó el discurso **con metáforas**
		loc / *locative*	Clara y Juan residen **en Barcelona**
		tem / *theme*	Clara recibió un regalo **de Juan**
	arg3	**ein** / *initial state*	El semáforo pasó **de verde** a rojo
		ins / *instrumental*	Clara se cepilla **con un cepillo azul**
		loc / *locative*	Nos alertaron **en un comunicado**
		ori / *origin*	Juan ha regresado **de las vacaciones**
	arg4	**des** / *destination*	Clara vino **a casa** ayer
		efi / *final state*	El semáforo pasó de verde **a rojo**
	argM	**adv** / *non-specific adjunct*	Juan vive **con su hermano**
		atr / *attribute*	El día amaneció **cubierto de niebla**
		cau / *cause*	Rompió la foto **por los malos recuerdos**
		ext / *extension*	Clara asumió **cinco años más** el cargo
		fin / *goal*	Lo hizo **para poder dormir a gusto**
		ins / *instrumental*	Ganó la carrera **con una bicicleta nueva**
		loc / *locative*	A Clara le gusta leer **en el jardín**
		mnr / *manner*	Juan lo hace todo **a su manera**
		tmp / *temporal*	Prefiere comer **a las 2**
ao / *sentence adjunct*	*none*	*none*	**Según ella**, Juan se lo merece
et / *textual element*	*none*	*none*	**Y** ¿qué dice él?
mod / *verbal modifier*	*none*	*none*	**No** quiso venir
impers / *impersonality marker*	*none*	*none*	**Se** trata de llegar a tiempo
pass / *passive marker*	*none*	*none*	Los importes **se** calculan en dólares

Table 4: <func>, <arg> and <tem> labels in AnCora 2.0 corpora

From Taulé et. al (2011) p.23

Annex B: An instance from our corpus

```
<tweet id="2929707725">
        <sentence>
                <v verb="es" token_id="19" sense="1">
                        <sn arg="arg2">
                                <word word="una" token_id="20"/>
                                <word word="eina" token_id="21"/>
                                <word word="gratuita" token_id="22"/>
                                <word word="i" token_id="23"/>
                                <word word="individualitzada"
                                   token_id="24"/>
                                <word word="per" token_id="25"/>
                                <word word="rebre" token_id="26"/>
                                <word word="suport" token_id="27"/>
                                <word word="juridic" token_id="28"/>
                        </sn>
                </v>
                <v verb="rebre" token_id="26" sense="1">
                        <sn arg="arg1">
                                <word word="suport" token_id="27"/>
                                <word word="juridic" token_id="28"/>
                        </sn>
                </v>
        </sentence>
        <sentence>
                <v verb="cal" token_id="37" sense="1">
                        <sn arg="arg1">
                                <word word="omplir" token_id="38"/>
                                <word word="el" token_id="39"/>
                                <word word="formulariURLURL" token_id
                                   ="40"/>
                        </sn>
                </v>
                <v verb="omplir" token_id="38" sense="1">
                        <sn arg="arg1">
                                <word word="el" token_id="39"/>
                                <word word="formulariURLURL" token_id
                                   ="40"/>
                        </sn>
                </v>
        </sentence>
</tweet>
```

Annex C: New entries to AnCora lexicon

Verb.deixar.11.default

Propbank	VerbNet	FrameNet	WordNet
Die	Disappearance-48.2 (die)	Death	Die (1)

Arguments

Function	Argument	Theme
suj	arg0	agt
cd	arg1	pat
ci	arg2	ben

Example:

Tweet: 820730191974662145

La Germandat Sant Ecce-Homo, està avui de dol. Ens **ha deixat** el soci i amic Albert Vallví i Navarro. Gran persona, gran tarragoní, amant de la Setmana Santa, va ser el Banderer l'any 2014.

Verb.vetllar.2.default

Propbank	VerbNet	FrameNet	WordNet
safeguard.01			

Arguments

Function	Argument	Theme
suj	arg0	agt
creg	arg1	tem
cc	argM	loc
cc	argM	tmp

Example:

Tweet: 998458326039126016

Som sostenibles!!!

A l'USER cada dia **vetllen** pel planeta.

Reciclen, cuiden de l'hort i, des de totes les àrees, treballen per a fer un món millor.

#thereisnoplanetB #sostenibilitat #reciclatge #ecoFEP #somesperança #ecofep #ecospes #escolessostenibles #mediambient URL

Verb.escriure.2.default

Propbank	VerbNet	FrameNet	WordNet
write.01	scribble-25.2	Text_creation	write (7)

Arguments

Function	Argument	Theme
suj	arg0	agt
cd	arg1	pat
ci	arg2	ben

Example:

Tweet: 22683937

«El llenguatge és laventura més gran de la meva vida, i la cerca de paraules és una cosa que em produeix molt de plaer».

Cita de Deborah Levy, **escriu** USER a USER

#AmbNosaltres8M URLURL

Verb.gaudir.3.default

Propbank	VerbNet	FrameNet	WordNet
enjoy.01	admire-31.2	Experiencer_subj	enjoy (1) enjoy (3) enjoy (5)

Arguments

Function	Argument	Theme
suj	arg0	agt
creg	arg1	tem
cc	argM	tmp
cc	argM	mnr

Example:

Tweet: 297327329

Amb l'arribada del bon temps **gaudeix** de la millor gastronomia vora el mar a qualsevol dels nostres 45 ports esportius. Des de la #CostaDaurada - Terres de l'Ebre, passant per la #CostaBarcelona i fins a la #CostaBrava

#portsdecatalunya #Catalunya #VitaminaBlava URL